Statistical Inference for Networks

4th Lehmann Symposium, Rice University, May 2011

Peter Bickel

Statistics Dept. UC Berkeley

(Joint work with Aiyou Chen, *Google*, E. Levina, *U. Mich*, S.

Bhattacharyya, *UC Berkeley*)

Outline

1. Networks: Examples
2. Descriptive statistics
3. Statistical issues and selected models
4. A nonparametric model for infinite networks and asymptotic theory
5. Statistical fitting approaches
   a) 'Moments'
   b) Pseudo likelihood
   c) Estimation of w

Example: Social Networks

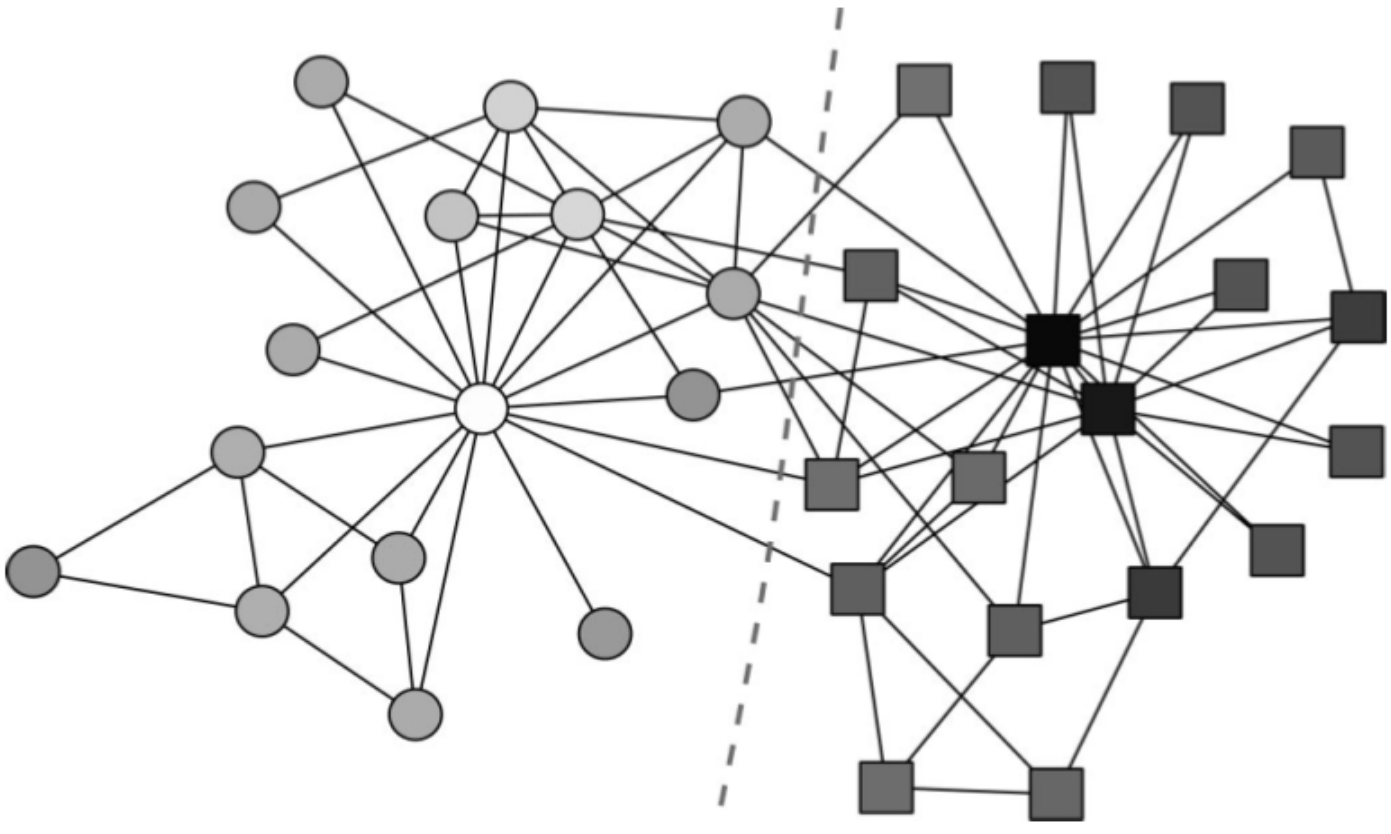


Figure: Karate Club (Newman, PNAS 2006)

Example: Social Networks

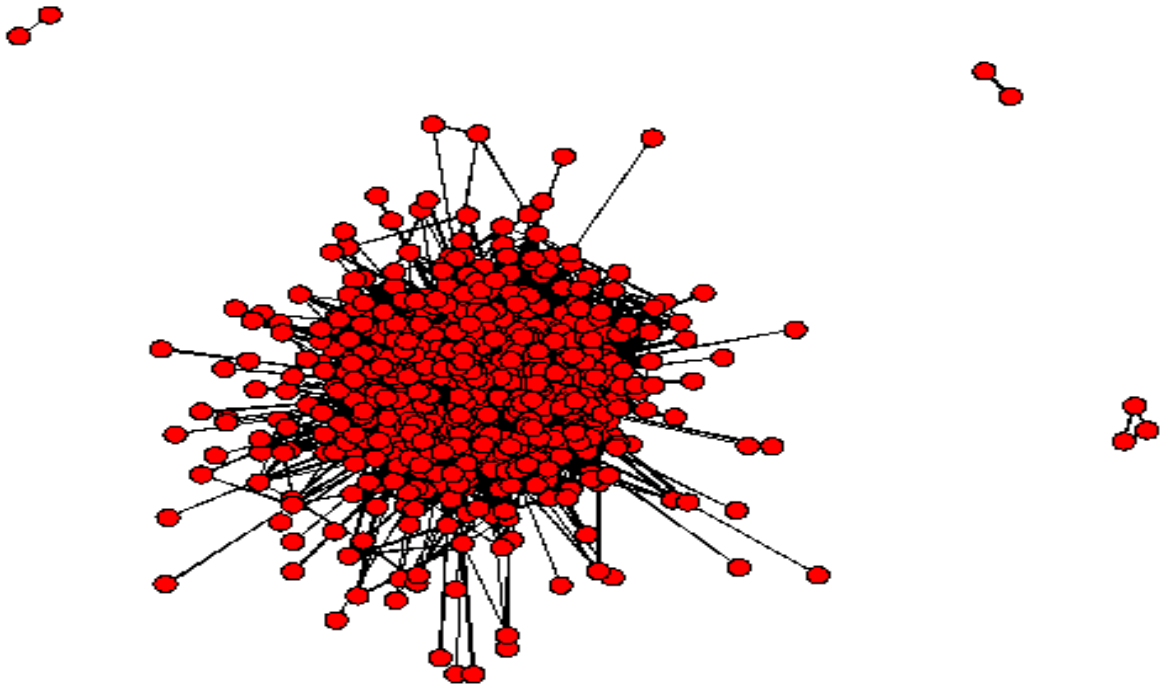


Figure: Facebook Network for Caltech with 769 nodes and average degree 43.

References

1. M.E.J. Newman (2010) Networks: An introduction. Oxford
2. Fan Chung, Linyuan Lu (2004) Complex graphs and networks. CBMS # 107 AMS
3. Eric D. Kolaczyk (2009) Statistical Analysis of Network Data
4. Bela Bollobas, Svante Janson, Oliver Riordan (2007) The Phase Transition in Random Graphs. Random Structures and Algorithms, 31 (1) 3-122
5. B. and A. Chen (2009) A nonparametric view of network models and Newman-Girvan and other modularities, PNAS
6. David Easley and Jon Kleinberg (2010) Networks, crowds and markets: Reasoning about a highly connected world. Cambridge University Press

A Mathematical Formulation

- $G = (V, E)$: undirected graph
- $\{1, \cdots, n\}$: Arbitrarily labeled vertices
- A : adjacency matrix
- $A_{ij} = 1$ if edge between i and j (relationship)
- $A_{ij} = 0$ otherwise
- $D_i = \sum_{j=1}^{n} A_{ij} =$ Degree of vertex i.

Descriptive Statistics

(Newman, Networks, 2010)

- Degree of vertex, Average degree of graph, $D_i = \sum_j A_{ij}$, $\overline{D}$
- # and size of connected components
- Geodesic distance
- Homophily $:= \frac{\#\text{ of }\Delta\text{'s}}{\#\text{ of }\Delta\text{'s} + \#\text{ of }V's}$.
- etc

Implications of Mathematical Description

- Undirected: Relations to or from not distinguished.
- Arbitrary labels: individual, geographical information not used. But will touch on covariates.

Stochastic Models

The Erdős-Rényi Model

- Probability distributions on graphs of n vertices.
- P on {Symmetric $n \times n$ matrices of 0's and 1's}.
- E-R (modified): place edges independently with probability λ/n ($\binom{n}{2}$ Bernoulli trials).

  $\lambda \approx E$(ave degree)

Nonparametric Asymptotic Model for Unlabeled Graphs

Given: P on ∞ graphs

Aldous/Hoover (1983)

$$\mathcal{L}(A_{ij} : i, j \geq 1\} \quad = \quad \mathcal{L}(A_{\pi_i, \pi_j} : i, j \geq 1),$$

for all permutations $\pi \iff$

$$\exists \quad g : [0,1]^4 \to \{0,1\} \text{ such that } A_{ij} \quad = \quad g(\alpha, \xi_i, \xi_j, \eta_{ij}),$$

where

α, ξ_i, η_{ij}, all i, $j \geq i$, i.i.d. $\mathcal{U}(0,1)$, $g(\alpha, u, v, w) = g(\alpha, v, u, w)$,

$\eta_{ij} = \eta_{ji}$.

Block Models (Holland, Laskey and Leinhardt 1983)

Probability model:

- Community label: $\mathbf{c} = (c_1, \cdots, c_n)$ i.i.d. multinomial $(\pi_1, \cdots, \pi_K) \equiv K$ "communities".
- Relation:

$$\mathbb{P}(A_{ij} = 1 | c_i = a, c_j = b) \quad = \quad P_{ab}.$$

- A_{ij} conditionally independent

$$\mathbb{P}(A_{ij} = 0) \quad = \quad 1 - \sum_{1 \leq a,b \leq K} \pi_a \pi_b P_{ab}.$$

- $K = 1$: E-R model.

Ergodic Models

$\mathcal{L}$ is an ergodic probability iff for g with $g(u, v, w) = g(v, u, w)$ $\forall(u, v, w)$,

$$A_{ij} = g(\xi_i, \xi_j, \eta_{ij}).$$

$\mathcal{L}$ is determined by

$$h(u, v) \equiv \mathbb{P}(A_{ij} = 1 | \xi_i = u, \xi_j = v),\ h(u, v) = h(v, u).$$

Notes:

1. K-block models and many other special cases
2. Model (also referred to as threshhold models) also suggested by Diaconis, Janson (2008)
3. More general models (Bollobás, Riordan & Janson (2007))

"Parametrization" of NP Model

- h is not uniquely defined.
- $h\big(\varphi(u), \varphi(v)\big)$, where φ is measure-preserving, gives same model.

  But, $h_{\mathrm{CAN}} =$ that $h(\cdot, \cdot)$ in equivalence class such that $P\left[A_{ij} = 1 | \xi_i = z\right] = \int_0^1 h_{\mathrm{CAN}}(z, v) dv \equiv \tau(z)$ with $\tau(\cdot)$ monotone increasing characterizes uniquely.
- ξ_i could be replaced by any continuous variables or vectors - but there is no natural unique representation.

Examples of models

i) Block models: on block of sizes π_a, π_b

$$h_{CAN}(u, v) \quad = \quad F_{ab}$$

ii) Power law: $$w(u, v) = a(u)a(v)$$

$$a(u) \quad \sim \quad (1-u)^{-\alpha} \text{ as } u \uparrow 1$$

iii) Dynamically defined model (preferential attachment):

$$w(u, v) = a(u)1(u \leq v) + a(v)1(u > v)$$

New vertex attaches to random old vertex and neighbors (not Hilbert-Schmidt)

$$a_{CAN}(u) = (1-u)^{-1} + \tau(u), \; a_{CAN}(u) = (1-u)^{-1} - \log(u(1-u))$$

Questions

i) Community identification and block models

ii) Checking "nonparametrically" with p "moments" whether 2 graphs are same (permutation tests used in social science literature for "block models", e.g., Wasserman and Faust, 1994).

iii) Link prediction: predicting relations to unobserved vertices on the basis of an observed graph.

iv) Model selection for hierarchies (block models).

v) Error bars on descriptive statistics.

vi) Linking graph features with covariates.

Asymptotic Approximation

- $h_n(u,v) = \rho_n w_n(u,v)$
- $\rho_n = \mathbb{P}[\text{Edge}]$
- $w(u,v)dudv = \mathbb{P}\left[\xi_1 \in [u, u+du], \xi_2 \in [v, v+dv] | \text{Edge}\right]$
- $w_n(u,v) = \min\left\{w(u,v), \rho_n^{-1}\right\}$
- Average Degree $= \frac{E(D_+)}{n} \equiv \lambda_n \equiv \rho_n(n-1)$.

Nonparametric Theory: The Operator

Corresponding to $w_{\mathrm{CAN}} \in L_2(0,1)$ there is operator:

$T : L_2(0,1) \to L_2(0,1)$

$Tf(\cdot) = \int_0^1 f(v) w(\cdot, v) dv$

T- Hermitian

Note: $\tau(\cdot) = T(\mathbf{1})(\cdot)$.

Nonparametric Theory

Let F and $\hat{F}$ be the distribution and empirical distribution of $\tau(\xi) \equiv T(\mathbf{1})(\xi)$ where ξ has a $U(0,1)$ distribution. Let $\rho = \lambda/n$.

Theorem 1

If $\lambda \to \infty$, then

$$\frac{1}{n}\sum_{i=1}^{n} E\left(D_i/\overline{D} - T(1)(\xi_i)\right)^2 \quad = \quad O(\lambda^{-1})$$

This implies, $\hat{F} \Rightarrow F$ in probability.

Identifiability of NP Model

Theorem 2

The joint distribution $(T(1)(\xi), T^2(1)(\xi), ..., T^m(1)(\xi), ...)$ where $\xi \sim U(0,1)$ determines P

Idea of proof: identify the eigen-structure of T.

Theorem 3

If T corresponds to a K-block model, then, the marginal distributions,

$$\left\{ T^k(1)(\xi) : k = 1, ..., K \right\}$$

determine (π, W) uniquely provided that the vectors π, $W\pi$, ..., $W^{K-1}\pi$ are linearly independent.

Methods of Estimation – Method of "Moments"

(k, ℓ)-wheel

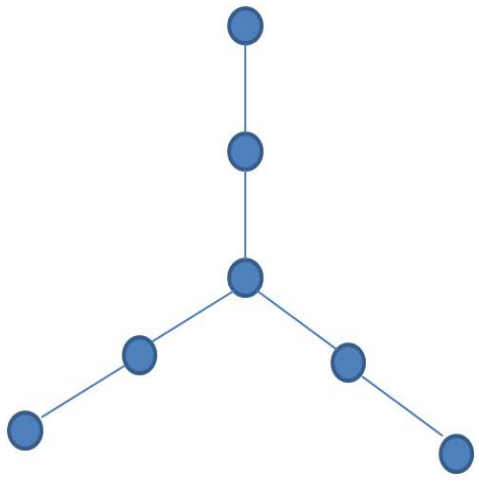
i) A "hub" vertex

ii) l spokes from hub

iii) Each spoke has k connected vertices.

Total # of vertices (order): $k\ell + 1$. Total # of edges (size): $k\ell$.

Eg: a (2,3)-wheel



"Moments"

- For $R \subset \{(i,j) \ : \ 1 \leq i < j \leq n\}$, identify R as a graph with vertex set $V(R) = \{i \ : \ (i,j) \text{ or } (j,i) \in R \text{ for some } j\}$ and $E(R) = R$.
- Let $G_n(R)$ be the subgraph induced by R in graph G_n.
- Define,

$$\begin{aligned} Q(R) &= P(A_{ij} = 1, \text{ all } (i,j) \in R) \\ P(R) &= P(E(G_n(R)) = R) \end{aligned}$$

- We can estimate $P(R)$ and $Q(R)$ in a graph G_n by

$$\begin{aligned} \hat{P}(R) &\equiv \frac{1}{\binom{n}{p} N(R)} \sum \left[\mathbf{1}(G \sim R : \ G \subset G_n), \ P(R) = E\hat{P}(R) \right] \\ N(R) &\equiv |\{G \subset G_n : \ G \sim R\}| \\ \hat{Q}(R) &\equiv \sum \{\hat{P}(S) : S \supset R\}, \ Q(R) = E\hat{Q}(R) \end{aligned}$$

Estimates of P and Q

Suppose $|R| = p$ fixed, $\rho_n \to 0$. Let $\mathbb{P}(h_n(\xi_1, \xi_2) > \rho) = o(n^{-1})$.

Then, define,

- $\tilde{P}(R) = \rho_n^{-p} P(R) = \tilde{Q}(R) + O(\lambda_n/n)$.
- $\tilde{Q}(R) = \rho_n^{-p} Q(R) \to E\left[\prod_{(i,j)\in R} w_n(\xi_i, \xi_j)\right]$.
- $\hat{\tilde{P}}(R) = \left(\frac{\overline{D}}{n}\right)^{-p} \hat{P}(R)$.
- $\hat{\tilde{Q}}(R) = \left(\frac{\overline{D}}{n}\right)^{-p} \hat{Q}(R)$.

Moment Convergence Theorem ($\lambda \to \infty$ and $\lambda = O(1)$)

Theorem 4

a) Suppose R is acyclic, and $\lambda \to \infty$.

$\sqrt{n}(\hat{\tilde{P}}(R) - \tilde{P}(R)) \Rightarrow \mathcal{N}(0, \sigma^2(R, P))$

and multivariate normality holds for $R_1, \cdots, R_k$ acyclic.

b) If $\lambda = O(1)$, a) continues to hold except that σ^2 depends on λ as well as R.

c) Even if R is not acyclic, the same conclusions apply to $\hat{\tilde{P}}$ and $\hat{\tilde{Q}}$ if $\lambda \geq n^{1-2/p}$.

Connection With Wheels

Lemma 1

Let G be a random graph generated according to P, $|V(G)| = k\ell + 1$. Then if R is a (k, ℓ)-wheel,

$$\begin{aligned} Q(R) &= E[T^k(1)(\xi_1)]^\ell \\ N(R) &= \frac{(kl+1)!}{\ell!} \\ P(R) &= Q(R) + O(\lambda/n) \end{aligned}$$

Difficulties

Even for sparse models

(i) Empirical moments of trees are hard to compute.

(ii) Empirical moments of small size converge reasonably even in sparse case, but block model parameters expressed as nonlinear function of moments not so well.

Extensions: Generalized Wheels

A $(\mathbf{k}, \mathbf{l})$-wheel, where $\mathbf{k} = (k_1, \ldots, k_t)$, $\mathbf{l} = (l_1, \ldots, l_t)$ are vectors and the k_j's, l_j's are distinct integers, is the union $R_1 \cup \cdots \cup R_t$, where R_j is a (k_j, l_j)-wheel, sharing a common hub but all their spokes are disjoint.

- Trees are examples of $(\mathbf{k}, \mathbf{l})$-wheels.
- Their limits yield cross-moments of $\left(T(\xi), T^2(\xi), \ldots\right)$.
- So, in principle, we can estimate parameters of block model, using the $(\mathbf{k}, \mathbf{l})$-wheels.
- Using $(\mathbf{k}, \mathbf{l})$-wheels, we can estimate the parameters of models approximating NP model.

Method of fitting: Pseudo likelihood

(Combining ideas of Besag (1974) and Newman & Leicht (2007))

Partition n into K communities of equal size

$S_1 = \{1, \cdots, m\}$,

$S_2 = \{m+1, \cdots, 2m\}$,

$\ldots$

$m = n/K$.

For each i: $b_{ik} = \sum\{A_{ij} : j \in S_k\}$.

a) Given c,

$$b_{ik} \sim \sum_{l \in S_k} \epsilon_{lk}$$

$$\epsilon_{lk} \text{ independent Bernoulli } (F_{c_i c_l})$$

$$b_{ik} \approx \text{ independent } Poiss(\lambda_{c_i,k}),$$

where $\lambda_{c_i,k} = n \sum_{s=1}^{K} r_{ks} F_{c_i s}$ and $r_{ks} = \frac{1}{n} \sum_{i \in S_k} 1(c_i = s)$.

b) Given $d_i = \sum_{k=1}^{K} b_{ik}$,

$$\{b_{ik} : k = 1, \cdots, K\} \sim \mathcal{M}(d_i, \{\theta_{c_i,k}\})$$

where $\theta_{ak} = \lambda_{ak} / \sum_{l=1}^{K} \lambda_{al}$, $k = 1, \cdots, K$.

Pseudo likelihood (cont)

Unconditionally on c:

a) $b_i \equiv \{b_{ik} : k = 1, \cdots, K\} \approx \sum_{j=1}^{K} \pi_j Poiss(\lambda_{jk})$

b) $\{b_{ik} : k = 1, \cdots, K\} \sim \sum_{j=1}^{K} \pi_j \mathcal{M}(d_i, \{\theta_{jk}\})$

Pretend b_i independent to get pseudo LogLikelihood:

a) $\sum_{i=1}^{n} \ell_i(\pi, \Lambda, b_i)$

a) $\sum_{i=1}^{n} \ell_i(\pi, \theta, b_i)$

Can be solved by simple EM, $\hat{\pi}$, $\hat{\Lambda}$, $\hat{\theta}$.

Theorem 5

Under appropriate identifiability conditions,

a) $\hat{\Lambda}, \hat{\theta}$ are consistent if $\frac{n^2\rho}{\log n} \to \infty$;

b) $\hat{\Lambda}, \hat{\theta}$ are $\sqrt{n}$ consistent if $n\rho = O(1)$.

Example: the Karate Club data ($K = 2$) (Zachary, 1977)

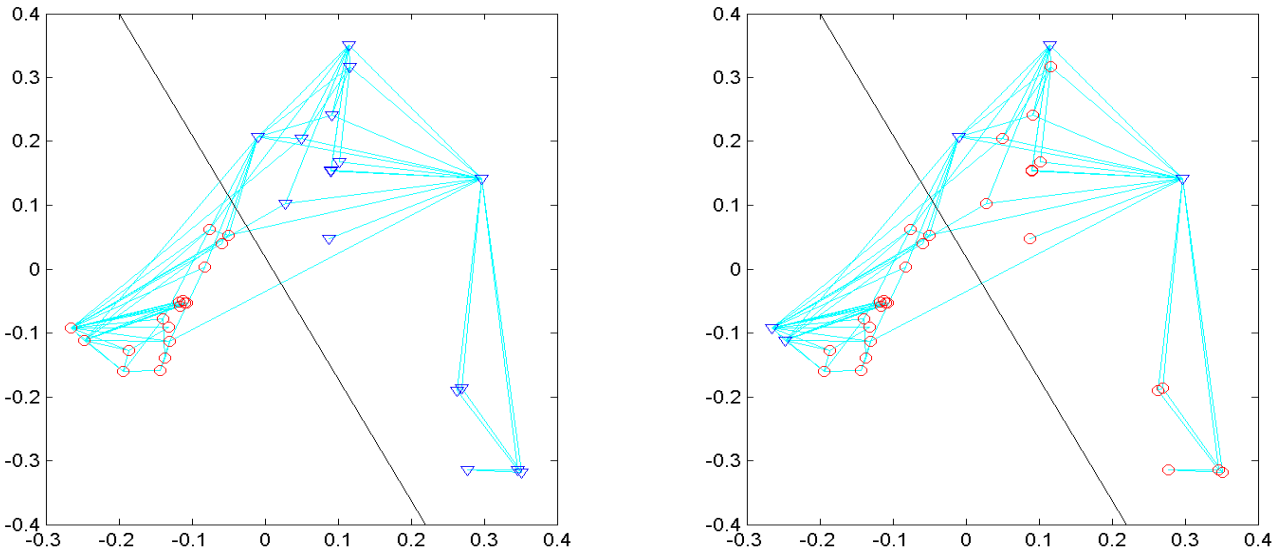


Figure: Left: conditional PL (correct classification), Right: unconditional PL (central nodes)

Advantages and Disadvantages of PL

1) PL a) is best for block models

2) PL has little theoretical justification.

3) PL also scales badly.

Can One Fit Nonparametric Model?

- Even parametric models are difficult to fit. We have seen that even for simple parametric models such as block models, the efficient estimation of the parameters is not easy.
- But still many of the parametric models are not good enough representation of the naturally occurring graphs. The empirical and theoretical vulnerability of Exponential Random Graph Models have been pointed out by Chatterjee and Diaconis (2010) and Bhamidi et. al. (2008).
- However, K block models seem to be attractive alternatives for modeling.

An Approach For Dense Models ($\lambda \to \infty$)

By Theorem 1(a), as $\lambda \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\tau(z_i) - \frac{D_i}{\bar{D}}\right)^2 = O\left(\frac{1}{\lambda}\right) \to 0 \tag{1}$$

here, $\tau(z) = T(\mathbf{1})(z)$.

Let

$$\hat{W}_n(u,v) = \int_0^u \int_0^v \frac{1}{n\bar{D}} \sum_{i,j} A_{ij}\mathbf{1}(\hat{\xi}_i \le s, \hat{\xi}_j \le t)dsdt$$

where $\hat{\xi}_i \equiv \hat{F}(\frac{D_i}{\bar{D}})$ and $\hat{F}$ is the empirical df of $\{\frac{D_i}{\bar{D}} : 1 \le i \le n\}$. Let

$$W_n(u,v) = \int_0^u \int_0^v \frac{1}{n\bar{D}} \sum_{i,j} A_{ij}\mathbf{1}(\xi_i \le s, \xi_j \le t)dsdt.$$

Theorem 6

Suppose that the conditions of Theorem 1 hold.

a) If $w(\cdot,\cdot)$ is bounded, and F, the df of $\tau(\xi_1)$, is Lipschitz and strictly increasing, then uniformly in (u, v),

$$|\hat{W}_n(u,v) - W_n(u,v)| = O_P\left(\frac{(\log\lambda)^{3/2}}{\lambda^{1/2}}\right).$$

Theorem 6 (cont)

b) If $\rho \to 0$ and $\tau(\xi_1)$ takes on only a finite number of values $t_1, \cdots, t_K$, then uniformly in (u, v),

$$|\hat{W}_n(u,v) - W_n(u,v)| \quad = \quad O_P(\lambda^{-1/2})|.$$

Moreover, if $W(u,v) = \int_0^1 \int_0^1 w(s,t)(u-s)_+(v-t)_+ dsdt$, then uniformly in (u, v),

$$|W_n(u,v) - W(u,v)| \quad = \quad O_P(\lambda^{-1/2})|.$$

Note:

$$\frac{\partial^4 W(u,v)}{(\partial u)^2(\partial v)^2} \quad = \quad w(u,v). \qquad (2)$$

An approach

a) Find smoothed empirical distribution function of $\frac{D_i}{\bar{D}}$,

$$\hat{F}(x) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left(\frac{D_i}{\bar{D}} \leq x\right)$$

b) Divide $[0,1]$ into intervals $I_1, \ldots, I_M$, such that, $I_j = [\frac{j-1}{M}, \frac{j}{M})$,

$$\begin{aligned}\hat{w}(u,v) \quad \equiv \quad & \frac{1}{\bar{D}} \sum_{a,b=1}^{M} \frac{1}{n^*} \mathbf{1}(u \in I_a)\mathbf{1}(v \in I_b) \\ & \times \left[\sum_{i,j=1}^{n} \mathbf{1}\left\{A_{ij} \ : \ \hat{F}\left(\frac{D_i}{\bar{D}}\right) \in I_a, \ \hat{F}\left(\frac{D_j}{\bar{D}}\right) \in I_b\right\}\right]\end{aligned}$$

where, $n^* = |I_a||I_b|$, if, $a \neq b$ and $n^* = (|I_a|(|I_a|-1))/2$, if, $a = b$.

Example: 2 Block Model

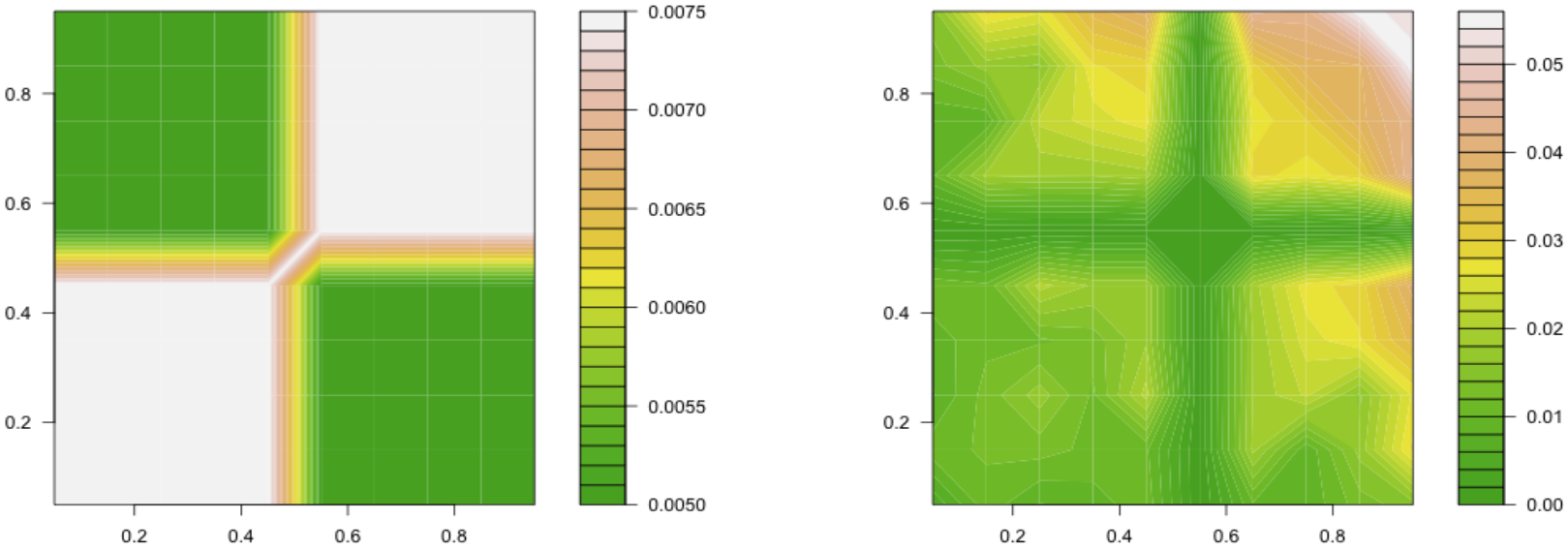


Figure: The LHS figre is the actual 2 block h function and RHS is the estimate of the h_{CAN} function.

Example: Facebook Caltech Network

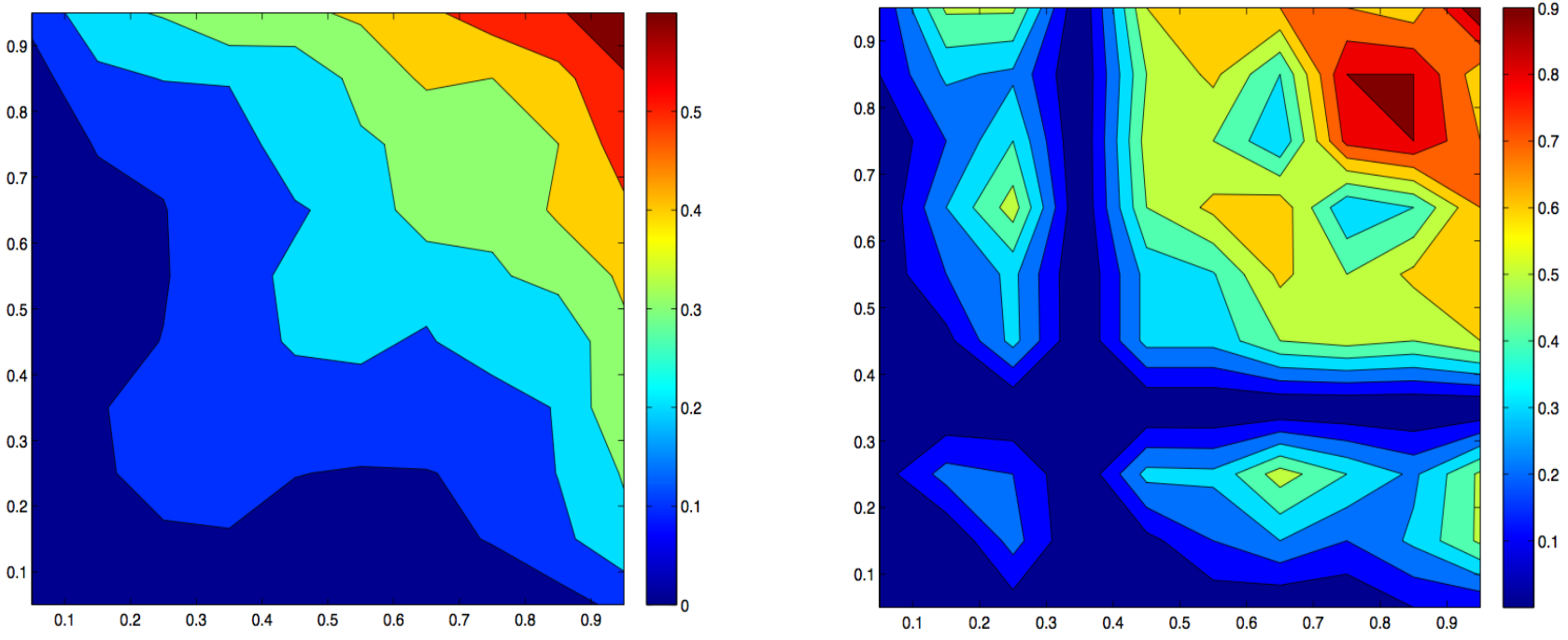


Figure: The LHS is estimate of h_{CAN} function for network of students of year 2008 and RHS is network of students of year 2008 residing in only 2 dorms. The proportions of classes in 2 distant modes are (0.3, 0.7) and (0.84, 0.16).

Why is the Result for Whole Network Uninstructive?

- $\xi \in U(0,1)$, w_{CAN} determine the probability uniquely but there are equivalent representation, which give very different results.
- $\xi \rightarrow$ degree suggest 'affinity', which is like 'linear' or first-order relation.
- We can now introduce higher-order relations, by making ξ a vector, that is, $\boldsymbol{(}\xi) = (\xi^{(1)}, \xi^{(2)})$, where, $\xi^{(1)}, \xi^{(2)} \sim U(0,1)$, $\xi_1 \perp \xi_2$.
- One way of forming $\xi^{(1)}, \xi^{(2)}$ is: let the binary representation of ξ is $\xi = (\xi_1, \xi_2, \xi_3, \xi_4, \ldots)$. Now define, $\xi^{(1)} = (\xi_1, \xi_3, \ldots)$ and $\xi^{(2)} = (\xi_2, \xi_4, \ldots)$.
- We know that, if $\xi \sim U(0,1)$, then, $(\xi^{(1)}, \xi^{(2)}) \sim U(0,1)^2$. Also, $\xi \rightarrow (\xi^{(1)}, \xi^{(2)})$ is 1-1 onto.

Example: 3 block Model

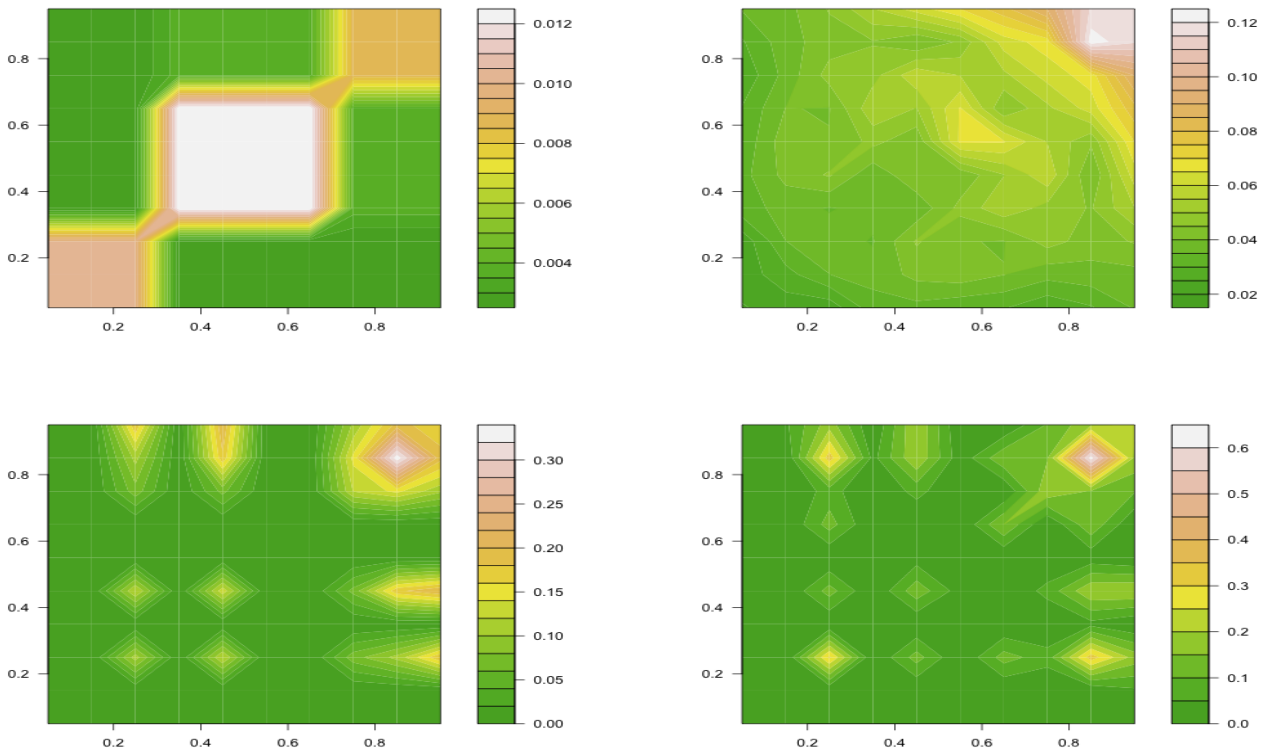


Figure: The top LHS figre is the actual 2 block h function and RHS is the estimate of the h_{CAN} function. The bottom LHS figure is the projection $\hat{h}_{CAN}(0.95,,0.95,)$ with two latent variables and bottom RHS figure is the sum of projections $\hat{h}_{CAN}(i,,i,)$ with two latent variables.

Example: Facebook Caltech Network

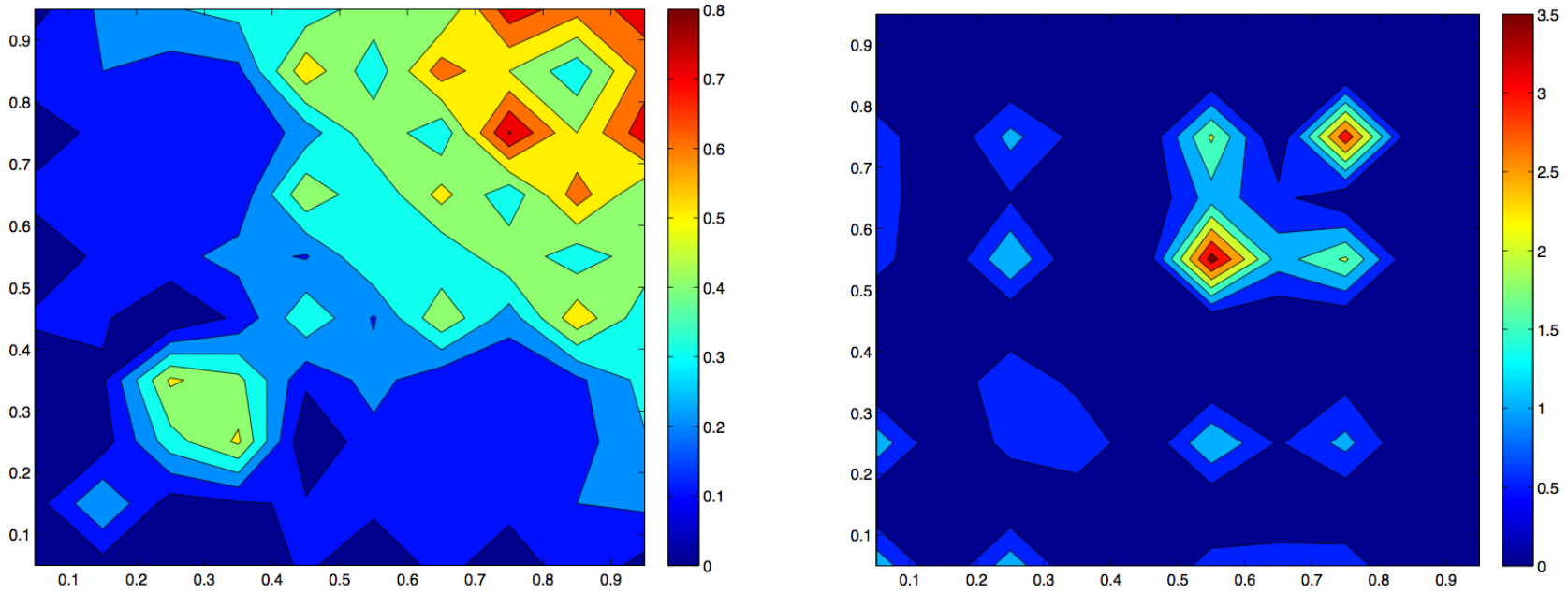


Figure: The LHS is estimate of h_{CAN} function for network of students of year 2008 residing in 3 dorms and RHS is sum of projections $\hat{h}_{CAN}(i,,i,)$ with two latent variables. The proportions of classes in 4 modes are (0.5, 0.13, 0.37), (0.67, 0.11, 0.22), (0.26, 0.66, 0.08), (0.32, 0.18, 0.5)

THANK YOU!

Examples of Social Networks

	Network	n	c	ℓ
Social	Film actors	449 913	113.43	3.48
	Company directors	7 673	14.44	4.60
	Math coauthorship	253 339	3.92	7.57
	Physics coauthorship	52 909	9.27	6.19
	Biology coauthorship	1 520 251	15.53	4.92
	Telephone call graph	47 000 000	3.16	
	Email messages	59 812	1.44	4.95
	Email address books	16 881	3.38	5.22
	Student dating	573	1.66	16.01
	Sexual contacts	2 810		
Information	WWW nd.edu	269 504	5.55	11.27
	WWW AltaVista	203 549 046	7.20	16.18
	Citation network	783 339	8.57	
	Roget's Thesaurus	1 022	4.99	4.87
	Word co-occurrence	460 902	66.96	
Technological	Internet	10 697	5.98	3.31
	Power grid	4 941	2.67	18.99
	Train routes	587	66.79	2.16
	Software packages	1 439	1.20	2.42
	Software classes	1 376	1.61	5.40
	Electronic circuits	24 097	4.34	11.05
	Peer-to-peer network	880	1.47	4.28
Biological	Metabolic network	765	9.64	2.56
	Protein interactions	2 115	2.12	6.80
	Marine food web	134	4.46	2.05
	Freshwater food web	92	10.84	1.90
	Neural network	307	7.68	3.97

Basic statistics: total number of vertices (n), mean degree (c), mean geodesic distance between connected vertex pairs (l)

Newman (2010) Networks: an introduction, Oxford